P3: Q&A

1. What were the objectives for this project?
Our team's goals were to analyze the psychological, physiological, social, environmental, and academic stressors experienced by students and develop a predictive model that can accurately forecast stress levels.

2. What steps did you take to achieve your objectives?
For our project, we started with data wrangling to clean and prepare the dataset, followed by EDA to understand trends and relationships. Afterward, we selected and trained our selected machine learning models to predict stress levels.

3. What dataset did you use for this project?
We used a dataset we found from Kaggle, consisting of 20 features impacting student stress levels. These stressors are categorized into psychological, physiological, social, environmental, and academic groups.

4. How did you clean and transform your dataset for this project?
We first converted object-type features into floats by removing symbols. We then assessed and confirmed that no missing values existed. Lastly, we made sure all features were numerical for the formatting and that the data types were compatible for further analysis and model training.

5. What methods did you use for exploratory data analysis (EDA)?
We built histograms to understand the distribution of each feature. We also constructed correlation matrices to understand the relationships between each feature and stress levels.

6. What insights did you extract from the histograms?
From the histograms, we were able to categorize features as either evenly distributed or skewed. Based on Fig. 1, 14 features are logged in as skewed, and 7 features are registered as evenly distributed. As the majority of our features are skewed, we noted to normalize these input variables for our model training process.

7. What insights did you extract from the correlation matrix?
Based on Fig. 2, we identified that bullying, future career concerns, and anxiety levels have high correlations with stress levels, indicating that as these factors increase, so do the stress levels. On the other hand, features like self-esteem, sleep quality, and academic performance show strong negative correlations, suggesting that higher values in these areas bring out lower stress levels.

Based on Fig. 3 and the color scaling of the matrices, we identified that psychological factors have the strongest influence on stress levels, while environmental factors are the least influential.

8. Which machine learning did you consider, and why?
We chose the Random Forest Classifier because it is known to effectively capture non-linear relationships between input and target variables. We also chose to use a Classifier rather than a Regressor as our target variable is categorical, even though the categories are numerically encoded.

We selected KNN because it performs well with numerical data. Since our features are already quantified using scales and ratings, we figured KNN would efficiently calculate distances between data points.

9. Could you summarize your model training process?
We trained our models using all features, adjusted hyperparameters via Grid Search, normalized features, and experimented with different train-test splits.

10. What were the results of your model training process?
Though the outputs of the two algorithms were close, we ultimately chose the Random Forest Classifier as our final model. Our Classifier model showed reliable performance with tuned hyperparameters, showing signs of balance between high accuracy and overfitting prevention.